-Oncology Standardization Challenge
-OHDSI Solution
-MSK Implementation

MITRE Community Presentation

Rimma Belenkaya, MSK
June 26, 2020
Introduction

Rimma Belenkaya

Head of Knowledge Management Team

Memorial Sloan Kettering Cancer Center (MSK)

MSK is a cancer treatment and research institution in New York City, founded in 1884 as the New York Cancer Hospital. It is the largest and oldest private cancer center in the world, and is one of 70 NCI-designated Comprehensive Cancer Centers.

Co-Lead of OHDSI Oncology Workgroup

OHDSI (formerly OMOP) stands for Observational Health Data Sciences and Informatics collaborative founded in 2009 with a base at Columbia University. OHDSI community conducts research on a global scale using standardized representations of health care data and reproducible standardized open source analytics.
The challenges of Real-World Data

What will it require?

- Data interoperability
- Standardised analytics
- Data network
- Strong community
OMOP Common Data Model (CDM) Version 6
OMOP Vocabulary

- Incorporate (=re-use), don’t invent

- Selected by
  - Comprehensiveness
  - Quality
  - Ontological principles: Concepts, life cycle, existing relationships to other

- Use Open Source, give credit

- Incorporate, disambiguate, map
Generating Reliable Evidence

At scale: repeatable replicable reproducible generalisable robust calibrated

---

Memorial Sloan Kettering Cancer Center
2015 – NYC-CDRN Success

- Standardized data of 7 mln patient from 6 major New York hospitals to OMOP
- Data are used for comparative effectiveness research and cohort selection for prospective studies
2016 - MSK Data Challenge

Lack of standardized data precludes or makes challenging

• **Cohort** generation for clinical studies

• **Data integration** to support analysis and research at scale

• **Data sharing** and **collaborative/network research**

• Ability to generate **reproducible evidence**

• Ability to **use existing medical knowledge** to drive clinical decision making and discovery
2016 - OHDSI Oncology WG is born

Michael Gurley, Northwestern
Christian Reich, IQVIA
Dmitry Dymshyts, Odysseus
Robert Miller, Tufts
Andrew Williams, Tufts
Rimma Belenkaya, MSK

Shilpa Ratwani, IQVIA, US
Asieh Golozar, Regeneron, US
Seng Chan You, Ajou University, South Korea
Eduard Korchmar, Odysseus, US
Vlad Korsik, Odysseus, US
Anastasios Siapos, IQVIA, Great Britain

Meera Patel, MSK, US
Shantha Bethusamy, MSK, US
Jeremy Warner, Vanderbilt University, US
Donna Rivera, NIH, US
Scott Campbell, University of Nebraska, US
RuiJun Chen, Columbia University, US

MISSION
Provide foundation for representing cancer data to support observational cancer research

Goals
- Extend OMOP CDM and Vocabulary for representing cancer data
- Develop ETL for tumor registry and EHR data
- Create use-case-driven algorithms for
  o identifying disease episodes, treatment pathways, and disease progression
  o identifying & characterizing cancer populations
  o predicting outcomes

https://github.com/OHDSI/OncologyWG/wiki#Oncology-WG
2016 – OHDSI Challenge: Granularity

Normal Condition
Most normal conditions are defined by three main dimensions implicitly, plus some extra attributes.

- Granulomatous infection
- Lung

Mycobacterium tuberculosis

Cancer
• Cause is not known, but morphology and topology are detailed and explicit.
• The many tumor attributes (modifiers) are also explicit and well defined.

- Carcinoma, NOS
- Breast, NOS

IIB: T2-N1-M0
- Stage
- Morphology
- Topology (site)
- HER2 status
- Positive

G3: High
- Grade
- Size
- Effected lymph nodes
- Distant metastases
- None

45 mm
4

Memorial Sloan Kettering Cancer Center
Challenges: Treatment Identification

Correct identification of cancer treatment regimens tends to be more complex compared to other disease modalities within observational data.

- Rarely treatment regimens are available in EMR
- Most cancer treatments are administered in chemotherapy regimens with complex dosing and scheduling in multiple cycles and are often combined with targeted therapies, immunotherapies, surgery or radiotherapy.
- None of these attributes follow standard definitions.
- Many regimens are personalized to the individual patient need, making a priori standardized definitions more complex.
Clinically and analytically relevant representation of cancer diagnoses, treatments, and outcomes requires data abstraction – Not readily available in the source data – Traditionally not supported in OMOP CDM
Challenges: Variety of Representations

Variety of Sources

- Cancer Registries with different and changing requirements
- Various EHR systems with no standard Oncology interfaces

Variety of Standards

- No single standard in any oncology domains
- Standards change along with the field advancements
Solving Granularity Challenge

Cancer Diagnosis Model in the OMOP Vocabulary

Added vocabularies:

- Carcinoma of Breast, NOS (8010/3-C50.9)
- Carcinoma, NOS
- Breast, NOS

Grade

- Grade I
- Grade II
- T-Cell

Tumor Size

- Numeric
- mm
- 001-988
  - 001
  - 988

Condition to schema

Schema to modifier

Memorial Sloan Kettering Cancer Center
Solving Granularity Challenge

Cancer diagnosis representation in the OMOP CDM

- Precoordinated concept of cancer Morphology + Site is stored in Condition_Occurrence
- Diagnostic modifiers are stored in Measurement and linked to the Condition_Occurrence record

Example of cancer diagnosis in the OMOP CDM

**Histology+Site** diagnosis in Condition_Occurrence

<table>
<thead>
<tr>
<th>condition_occurrence_id</th>
<th>123456789</th>
</tr>
</thead>
<tbody>
<tr>
<td>person_id</td>
<td>1</td>
</tr>
<tr>
<td>condition_concept_id</td>
<td>4116071</td>
</tr>
<tr>
<td>condition_start_datetime</td>
<td>June 9, 2019</td>
</tr>
<tr>
<td>condition_type_concept_id</td>
<td>32535</td>
</tr>
<tr>
<td>condition_source_value</td>
<td>8010/3-C50.9</td>
</tr>
<tr>
<td>condition_source_concept_id</td>
<td>44505310</td>
</tr>
</tbody>
</table>

**Grade** modifier in Measurement

<table>
<thead>
<tr>
<th>measurement_id</th>
<th>567890</th>
</tr>
</thead>
<tbody>
<tr>
<td>person_id</td>
<td>1</td>
</tr>
<tr>
<td>measurement_datetime</td>
<td>June 9, 2019</td>
</tr>
<tr>
<td>measurement_concept_id</td>
<td>35918640</td>
</tr>
<tr>
<td>measurement_date</td>
<td>June 9, 2019</td>
</tr>
<tr>
<td>value_as_concept_id</td>
<td>35922509</td>
</tr>
<tr>
<td>measurement_type_concept_id</td>
<td>32534</td>
</tr>
<tr>
<td>measurement_source_value</td>
<td>3844</td>
</tr>
<tr>
<td>measurement_source_concept_id</td>
<td>35918640</td>
</tr>
<tr>
<td>value_source_value</td>
<td>breast@3844@3</td>
</tr>
<tr>
<td>modifier_of_event_id</td>
<td>123456789</td>
</tr>
<tr>
<td>modifier_field_concept_id</td>
<td>1147127</td>
</tr>
</tbody>
</table>
Example of disease and treatment episodes in the Episode table

**‘First occurrence’-of-‘Carcinoma of breast’**

- episode_id: 12345
- person_id: 1
- episode_concept_id: 32528
- episode_start_datetime: June 9, 2019

**‘Treatment regimen’-of-‘Carboplatin and Paclitaxel’**

- episode_id: 12346
- person_id: 1
- episode_concept_id: 32531
- episode_start_datetime: July 9, 2019
- episode_parent_id: 12345
- episode_object_concept_id: 35804255

OMOP concept
- ‘First disease occurrence’
- SNOMED concept
- ‘Carcinoma of breast’
- OMOP concept
- ‘Tumor registry’

OMOP concept
- ‘Treatment Regimen’
- Foreign key to the disease Episode record
- HemOnc concept
- ‘Carboplatin and Paclitaxel’
- OMOP concept
- ‘Episode algorithmically derived from EHR’
2019 – OMOP Oncology Module Released

- **OMOP CDM**
  - Supports diagnostic attributes
  - Supports episodes of disease and treatment
  - Supports explicit linkages between episodes and lower level events

- **OMOP Vocabulary**
  - ICD-O-3: golden standard for encoding cancer histology and topology
  - Mappings between ICD-O-3 and SNOMED
  - NAACCR: Standard for data collection in US Tumor Registry
  - CAP Data Dictionary: Standard for Synoptic Pathology Reports
  - HemOnc: Ontology of chemotherapy regimens

- **ETL and Analytics**
  - Vocabulary-driven ETL for US Tumor Registry data
  - Algorithm for extraction of chemotherapy regimens from medication records

- **Research**
  - Network characterization study of bladder cancer
2020-Roadmap

Hypothetical Patient Journey

Diagnosis → Advanced disease diagnosis → Remission → Progression → Stable disease → Progression → Hospitalization

- Surgery/Radiation
- Biomarker evaluation
- 1L Therapy
- 2L Therapy
- 3L Therapy

Accomplished:
- Granular representation of cancer based on ICD-O and NAACCR
- Ingestion of tumor registry data using vocabulary-driven ETL
- Derivation of First Cancer Occurrence Episode from tumor registry
- Derivation of First Treatment Course Episode including Chemotherapy Regimens from tumor registry
- Derivation of Chemotherapy regimen from lower level medication

Under Development:
- Expansion of Drug Regimen Algorithm, improving precision of chemo regimen identification
- Genomic-CDM
- EHR into OMOP CDM
- Incorporation of CAP in the vocabulary
- Standardized vocabulary for representing diagnostic modifiers
- Development of models and vocabularies for disease progression, response, and other episodes

Future:
- Development of algorithms for identifying disease progression/response and other episodes
- Domain for Imaging
- Data quality checks for NAACCR ETL

Examples of questions that be addressed now:
- Characteristics of patients
- Treatment pattern, dosing, regimens and sequence
- Time to treatment discontinuation and time to next treatment
- Compare uptake of newer medications vs. older medications
- Number of medications taken daily by a cancer patient
- Overall survival from:
  - time of disease diagnosis
  - time of diagnosis with advanced stage
  - time of treatment initiation

Memorial Sloan Kettering Cancer Center
2020 – Implementation at MSK

• Conversion of EMR and Tumor Registry data – underway

• Conversion of COVID data – completed

• Integration of OMOP CDM with MSK Extract platform for research data - underway
2020 – mCODE -> OMOP

• **mCODE** model has been incorporated into MSK “pan-cancer” model

• Next: Crosswalk from MSK “pan-cancer” (mCODE) to
  1. OMOP vocabulary
  2. OMOP CDM